# PCI-SIG DRAFT ENGINEERING CHANGE NOTICE

| | |
|---|---|
| **TITLE:** | Native PCIe Enclosure Management |
| **DATE:** | May 18, 2017 |
| **AFFECTED DOCUMENT:** | PCI Express Base Spec. Rev. 3.x |
| **SPONSOR:** | Intel, Dell EMC, Microsemi, <Others?> (On behalf of NVMe-MI Working Group) |

## Part I

### 1. Summary of the Functional Changes

Defines mechanisms for simple storage enclosure management for NVMe SSDs, consistent with established capabilities in the storage ecosystem, with the first version of this capability defining a register interface for LED control.

This ECN defines a new PCI Express extended capability called Native PCIe Enclosure Management (NPEM).

### 2. Benefits as a Result of the Changes

Incumbent SAS/SATA storage ecosystem establishes a user interface model for indication of drive status (e.g., In a Failed Array, In a Critical Array, Rebuilding, Predicted to Fail, etc.)and integrates the logic inside a central SAS/SATA controller or an SAS Host Bus Adapter (HBA) that is separate from the drives.

In NVMe, the controller/HBA is part of each drive thus the notion of a separate central controller is eliminated. In typical whitebox server implementations for NVMe, enclosure function is inside a root port or a downstream switch port to which NVMe drive is connected. This takes the enclosure function outside of the NVMe subsystem and brings it under the purview of PCIe.

Through the mechanisms defined by this ECR, PCIe will provide an enclosure/LED capability/control model for NVMe drives/storage arrays that the incumbent storage technologies (e.g., SAS, SATA) already support.

## 3. Assessment of the Impact

NPEM will be an optional capability. There is no impact to hardware/software that do not support the new mechanisms. New storage hardware and software elements are expected to take advantage of the new capability defined herein.

## 4. Analysis of the Hardware Implications

Implementations that support NPEM capability will need to support new register interface and associated hardware logic to support the functionality/behaviors defined in this document. This capability can be supported inside an NVMe drive (PCIe endpoint), inside a root port, or inside a switch's downstream port.

This ECR only defines the register interface for software/firmware. The hardware details of how the LED control signals travel from root/switch downstream ports to the storage enclosure backplane, as well as the details of LED blink patterns are outside the scope of this ECR.

## 5. Analysis of the Software Implications

No changes are required to existing software, and existing software will not benefit from this feature.

New firmware/software is required to achieve the benefits associated with this capability.

It is intended that this capability can be controlled through non-vendor-specific software or through vendor-specific software, which would typically support a broader/richer feature set, and capability is defined to support both approaches.

## 6. Analysis of the C&I Test Implications

None anticipated other than the read/write register tests.

## Part II

**Detailed Description of the change**

*Modify Terms and Acronyms as shown:*

## Terms and Acronyms

…

NPEM

Native PCIe Enclosure Management

SSD

Solid State Drive

…

*Add new section:*

## 6.x. Native PCIe Enclosure Management

NPEM is an optional PCIe Extended Capability that provides mechanisms for enclosure management. This mechanism is designed to provide management for enclosures containing PCIe SSDs that is consistent with the established capabilities in the storage ecosystem.

This section defines the architectural aspects of the mechanism. The NPEM extended capability is defined in section 7.x.y.

An enclosure is any platform, box, rack, or set of boxes that contain one or more PCIe SSDs. The NPEM capability provides storage related enclosure control (e.g., status LED control) for a PCIe SSD. The NPEM capability may reside in a Downstream port, or an Endpoint (i.e., the PCIe SSD). Figure 6-x1 shows an example configuration with a single Downstream Port containing the NPEM capability and vendor specific logic to control the associated LEDs.

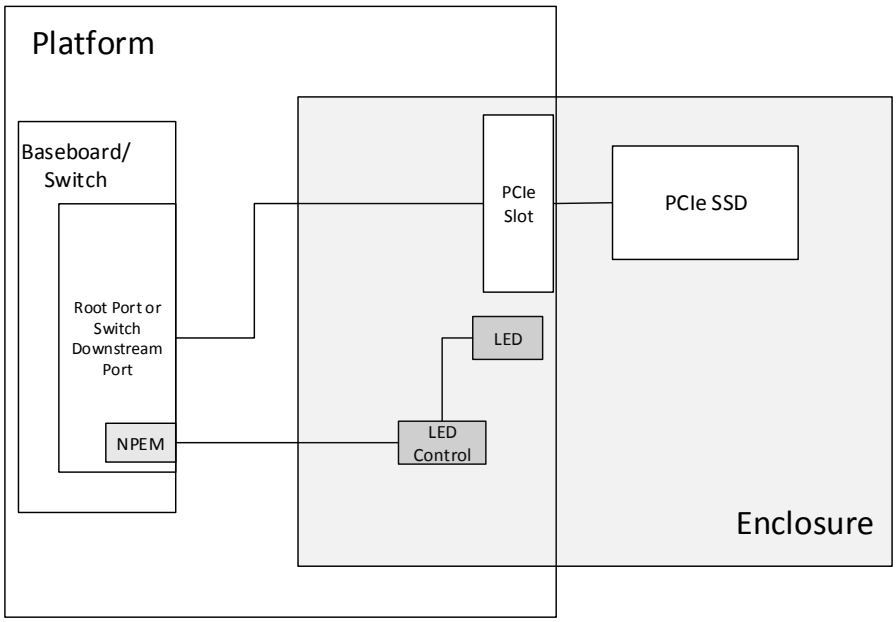


**Figure 6-x1 Figure: Example NPEM Configuration using a Downstream Port**

Figure 6-x2 shows an example configuration with the NPEM capability located in the Upstream Port (in this case, the SSD function).

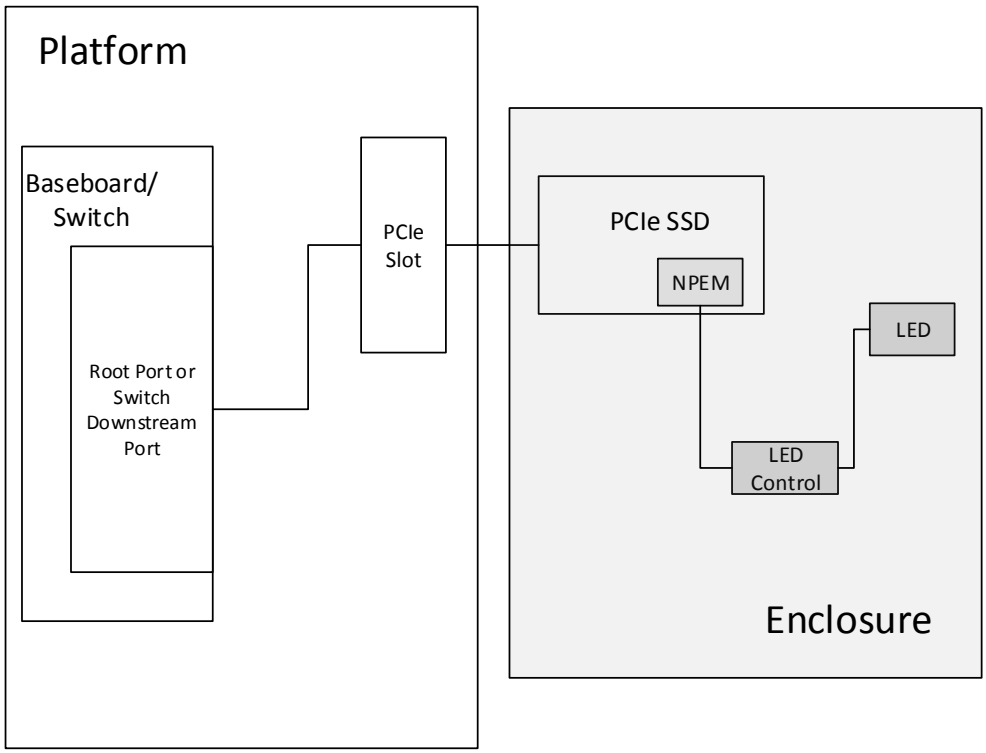


**Figure 6-x2: Example NPEM Configuration using an Upstream Port**

Software issues an NPEM command by writing to the NPEM Control register to change the indications associated with an SSD. NPEM Command is a single write to the NPEM Control register that changes the state of zero or more bits. NPEM indicates a successful

completion to software using the command completed mechanism. Figure 6-x3 shows the overall flow.

This specification defines the software interface provided by the NPEM capability. The Port to enclosure interface, enclosure, enclosure to LED interface, number of LEDs per SSD, and associated LED blink patterns are all outside the scope of this specification.

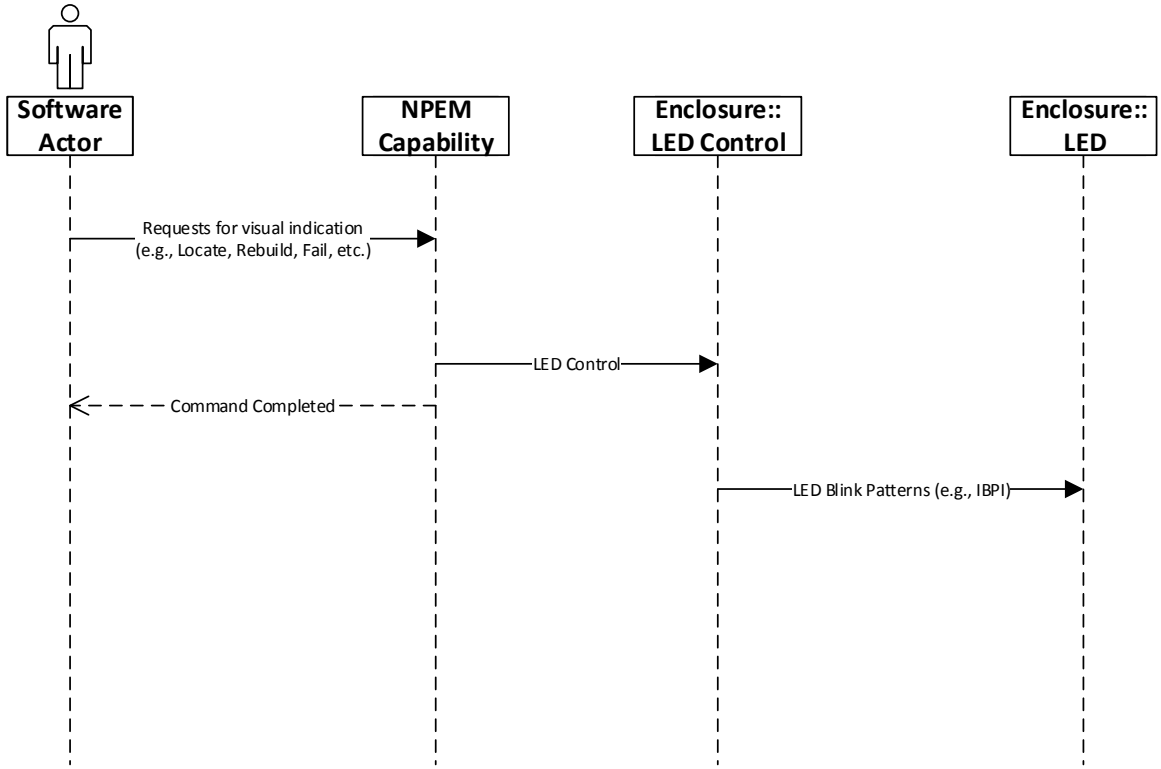


**Figure 6-x3: NPEM Command Flow**

NPEM provides a mechanism for system software to issue a reset to the LED control element within the enclosure by means of the NPEM Reset mechanism, which is independent of the PCIe link itself. The NPEM command completed mechanism also applies to NPEM Reset.

NPEM mechanism is orthogonal to PCIe hot plug mechanisms.A hardware relationship between them is vendor specific and outside the scope of this document.

Storage system admin or software controls the indications for various device states through the NPEM capability.

## Implementation Note

Table 6-x1 shows an example of NPEM states and a possible meaning that some enclosures may implement.

**Table 6-x1: NPEM States**

| NPEM State | Actor | Definition |
|---|---|---|

| | | |
|---|---|---|
| OK | System Admin or Storage Software | OK state may mean the drive is functioning normally. The NPEM OK state may implicitly mean that an SSD is present, powered on, and working normally as seen by the software. A more granular indication of drive not physically present or present but not powered up are both outside the scope of this specification. |
| Locate | System Admin | Locate state may mean the specific drive is being identified by an admin. |
| Fail | Storage Software | Fail state may mean the drive is not functioning properly |
| Rebuild | Storage Software | Rebuild state may mean this drive is part of a multi-drive storage volume/array that is rebuilding or reconstructing data from redundancy on to this specific drive. |
| PFA | Storage Software | PFA stands for Predicted Failure Analysis. This state may mean the drive is still functioning normally but predicted to fail soon. |
| Hot Spare | Storage Software | Hot Spare state may mean this drive is marked to be automatically used as a replacement for a failed drive and contents of the failed drive may be rebuilt on this drive. |
| In A Critical Array | Storage Software | In A Critical Array state may mean the drive is part of a multi-drive storage array and that array is degraded. |
| In A Failed Array | Storage Software | NPEM In A Failed Array state may mean the drive is part of a multi-drive storage array and that array is failed. |
| Invalid Device Type | Storage Software | Invalid Device Type state may mean the drive is not the right type for the connector (e.g., U.2 supports SAS and NVMe drives and this drive state indicates that a SAS drive is plugged into an NVMe slot). |
| Disabled | Storage Software | Disabled state may mean the drive in this slot is disabled. A removal of this drive from the slot may be safe. The power from this slot may be removed. |

## 7.x. Native PCIe Enclosure Management Extended Capability

The Native PCIe Enclosure Management Extended Capability (NPEM) is an optional extended capability that is permitted to be implemented by Root Ports, Switch Downstream Ports, and Endpoints.

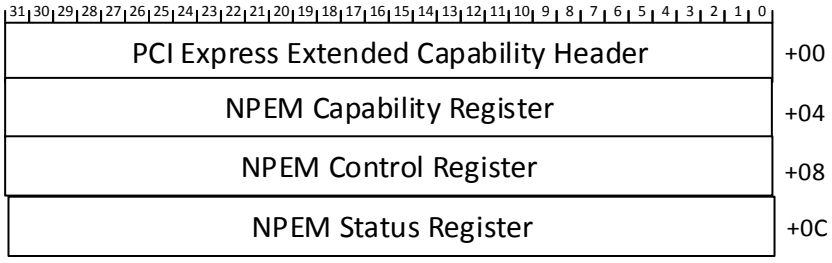


Figure 7-x1 NPEM Extended Capability

## 7.x.1. Native PCIe Enclosure Management Extended Capability Header (Offset 00h)

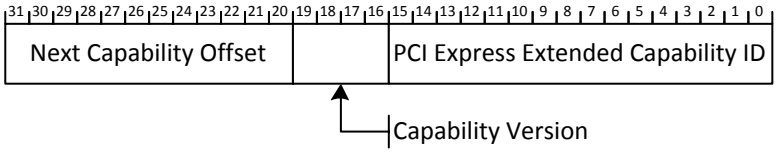


Figure 7-x1 NPEM Extended Capability Header

Table 7-x1 Native PCIe Enclosure Management Extended Capability Header

| Bit Location | Register Description | Attributes |
|---|---|---|
| 15:0 | **PCI Express Extended Capability ID** – This field is a PCI-SIG defined ID number that indicates the nature and format of the extended capability.<br><br>PCI Express Extended Capability ID for the NPEM Extended Capability is 0029h. | RO |
| 19:16 | **Capability Version** – This field is a PCI-SIG defined version number that indicates the version of the capability structure present.<br><br>Must be 1h for this version of the specification. | RO |
| 31:20 | **Next Capability Offset** – This field contains the offset to the next PCI Express Extended Capability structure or 000h if no other items exist in the linked list of capabilities. | RO |

## 7.x.2. Native PCIe Enclosure Management Capability Register (Offset 04h)

The NPEM Capability Register contains an overall NPEM capable bit and a bit map of states supported in the implementation. Implementations are required to support OK, Locate, Fail, and Rebuild states if NPEM capable bit is Set to 1b. All other states are optional.

Use of Enclosure Specific bits is outside the scope of this specification.

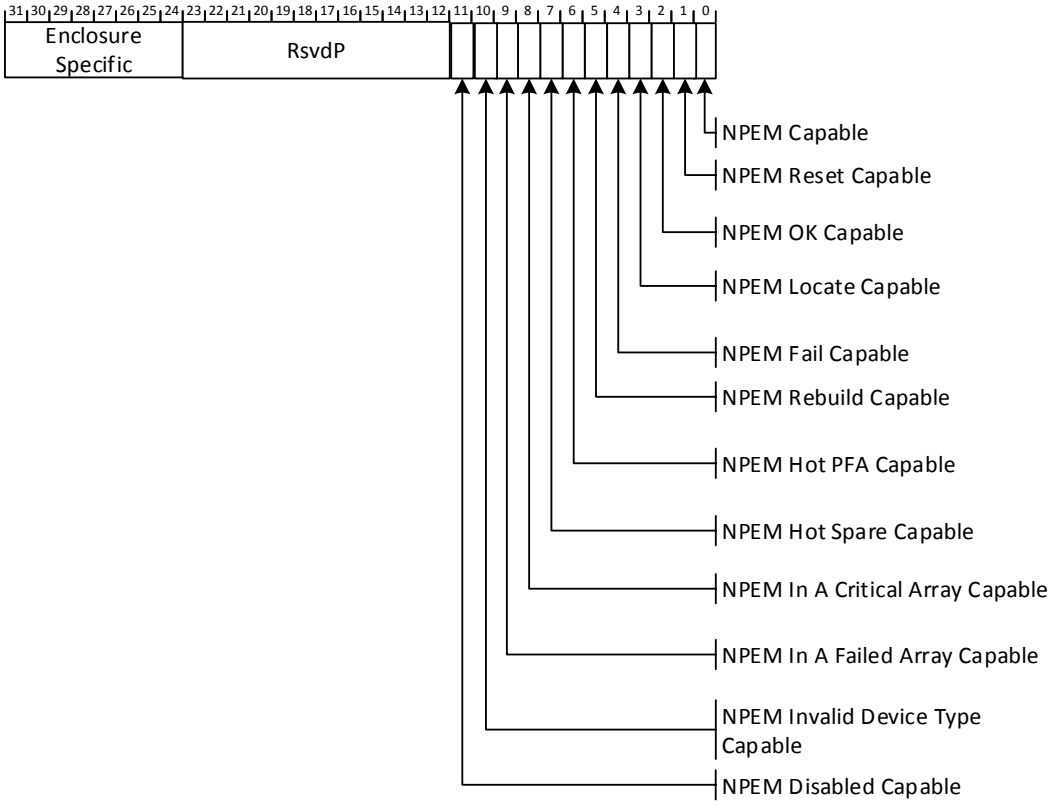


**Figure 7-x2 NPEM Capability Register**

5 **Table 7-x2 Native PCIe Enclosure Management Capability Register**

| Bit Location | Register Description | Attributes |
|---|---|---|
| 0 | **NPEM Capable-**When Set, this bit indicates that the enclosure has NPEM functionality. | HwInit |
| 1 | **NPEM Reset Capable-** A value of 1b indicates support for the optional NPEM Reset mechanism described in <Section 6.x.>. This capability is independently optional. | HwInit |
| 2 | **NPEM OK Capable-** When Set, this bit indicates that enclosure has the ability to indicate the NPEM OK state. This bit must be Set if NPEM Capable is also Set. | HwInit |
| 3 | **NPEM Locate Capable-** When Set, this bit indicates that enclosure has the ability to indicate the NPEM Locate state. This bit must be Set if NPEM Capable is also Set. | HwInit |
| 4 | **NPEM Fail Capable-** When Set, this bit indicates that enclosure has the ability to indicate the NPEM Fail state. This bit must be Set if NPEM Capable is also Set. | HwInit |

| | | |
|---|---|---|
| 5 | **NPEM Rebuild Capable-** When Set, this bit indicates that enclosure has the ability to indicate the NPEM Rebuild state. This bit must be Set if NPEM Capable is also Set. | HwInit |
| 6 | **NPEM PFA Capable -** When Set, this bit indicates that enclosure has the ability to indicate the NPEM PFA state. This capability is independently optional. | HwInit |
| 7 | **NPEM Hot Spare Capable -** When Set, this bit indicates that enclosure has the ability to indicate the NPEM Hot Spare state. This capability is independently optional. | HwInit |
| 8 | **NPEM In A Critical Array Capable -** When Set, this bit indicates that enclosure has the ability to indicate the NPEM In A Critical Array state. This capability is independently optional. | HwInit |
| 9 | **NPEM In A Failed Array Capable -** When Set, this bit indicates that enclosure has the ability to indicate the NPEM In A Failed Array state. This capability is independently optional. | HwInit |
| 10 | **NPEM Invalid Device Type Capable -** When Set, this bit indicates that enclosure has the ability to indicate the NPEM_Invalid_Device_Type state. This capability is independently optional. | HwInit |
| 11 | **NPEM Disabled Capable -** When Set, this bit indicates that enclosure has the ability to indicate the NPEM_Disabled state. This capability is independently optional. | HwInit |
| 23:12 | **Reserved** | RsvdP |
| 31:24 | **Enclosure Specific -** The definition of enclosure specific bits is outside the scope of this specification. | HwInit |

## 7.x.3. Native PCIe Enclosure Management Control Register (Offset 08h)

The NPEM Control Register contains an overall NPEM enable bit and a bit map of states that software controls.

Use of Enclosure Specific bits is outside the scope of this specification.

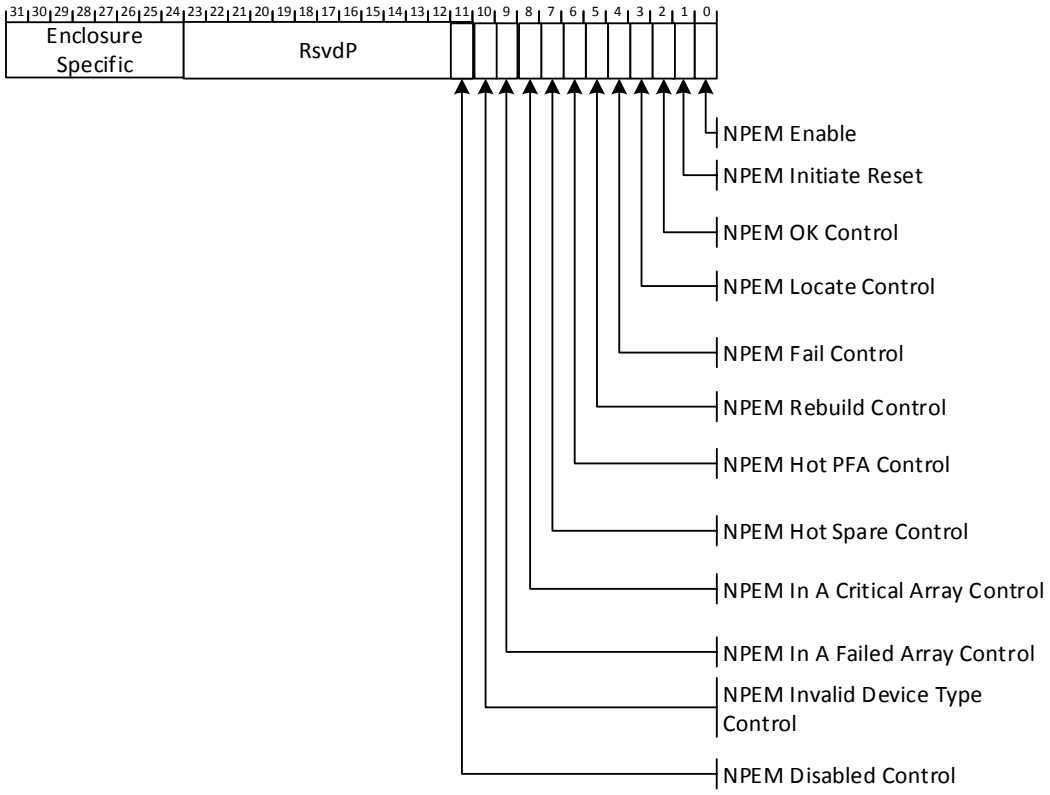


Figure 7-x3 NPEM Control Register

Table 7-x3 Native PCIe Enclosure Management Control Register

| Bit Location | Register Description | Attributes |
|---|---|---|
| 0 | **NPEM Enable-**When Set to 1b, this bit enables the NPEM capability. This bit Cleared to 0b disables the NPEM capability.<br><br>Default value of this bit is 0b<br><br>When enabled, this capability operates as defined in this specification. When disabled, the other bits in this capability have no effect and any associated indications are outside the scope of this specification.<br><br>Setting and clearing NPEM Enable is an NPEM command and will result in an NPEM command completion indication when finished. | RW |

| | | |
|---|---|---|
| 1 | **NPEM Initiate Reset –** If NPEM Reset Capability is 1b, then a write of 1b to this bit initiates NPEM Reset. If NPEM Reset Capability is 0b, then this bit is Reserved.<br><br>The value read by software from this bit must always be 0b.<br><br>Setting NPEM Reset is an NPEM command and will result in an NPEM command completion indication when finished. | RW/RsvdZ |
| 2 | **NPEM OK Control –** When Set to 1b, this bit specifies that the NPEM OK indication be turned ON. This bit Cleared to 0b specifies that the NPEM OK indication be turned OFF.<br><br>If NPEM OK Capable bit in NPEM Capability register is 0b, this bit is permitted to be read-only with a value of 0b.<br><br>Default value of this bit is 0b | RW |
| 3 | **NPEM Locate Control-** When Set to 1b, this bit specifies that the NPEM Locate indication be turned ON. This bit Cleared to 0b specifies that the NPEM Locate indication be turned OFF.<br><br>If NPEM Locate Capable bit in the NPEM Capability register is 0b, this bit is permitted to be read-only with a value of 0b.<br><br>Default value of this bit is 0b | RW |
| 4 | **NPEM Fail Control -** When Set to 1b, this bit specifies that the NPEM Fail indication be turned ON. This bit Cleared to 0b specifies that the NPEM Fail indication be turned OFF.<br><br>If NPEM Fail Capable bit in the NPEM Capability register is 0b, this bit is permitted to be read-only with a value of 0b.<br><br>Default value of this bit is 0b | RW |
| 5 | **NPEM Rebuild Control -** When Set to 1b, this bit specifies that the NPEM Rebuild indication be turned ON. This bit Cleared to 0b specifies that the NPEM Rebuild indication be turned OFF.<br><br>If NPEM Rebuild Capable bit in NPEM Capability register is 0b, this bit is permitted to be read-only with a value of 0b.<br><br>Default value of this bit is 0b | RW |
| 6 | **NPEM PFA Control -** When Set to 1b, this bit specifies that the NPEM PFA indication be turned ON. This bit Cleared to 0b specifies that the NPEM PFA indication be turned OFF.<br><br>If NPEM PFA Capable bit in NPEM Capability register is 0b, this bit is permitted to be read-only with a value of 0b.<br><br>Default value of this bit is 0b | RW |

| | | |
|---|---|---|
| 7 | **NPEM Hot Spare Control -** When Set to 1b, this bit specifies that the NPEM Hot Spare indication be turned ON. This bit Cleared to 0b specifies that the NPEM Hot Spare indication be turned OFF.<br><br>If NPEM Hot Spare Capable bit in NPEM Capability register is 0b, this bit is permitted to be read-only with a value of 0b.<br><br>Default value of this bit is 0b | RW |
| 8 | **NPEM In A Critical Array Control -** When Set to 1b, this bit specifies that the NPEM In A Critical Array indication be turned ON. This bit Cleared to 0b specifies that the NPEM In A Ciritical Array indication be turned OFF.<br><br>If NPEM In A Critical Array Capable bit in NPEM Capability register is 0b, this bit is permitted to be read-only with a value of 0b.<br><br>Default value of this bit is 0b | RW |
| 9 | **NPEM In A Failed Array Control -** When Set to 1b, this bit specifies that the NPEM In A Failed Array indication be turned ON. This bit Cleared to 0b specifies that the NPEM In A Failed Array indication be turned OFF.<br><br>If NPEM In A Failed Array Capable bit in NPEM Capability register is 0b, this bit is permitted to be read-only with a value of 0b.<br><br>Default value of this bit is 0b | RW |
| 10 | **NPEM Invalid Device Type Control -** When Set to 1b, this bit specifies that the NPEM Invaild Device Type indication be turned ON. This bit Cleared to 0b specifies that the NPEM Invalid Device Type indication be turned OFF.<br><br>If NPEM Invalid Device Type Capable bit in NPEM Capability register is 0b, this bit is permitted to be read-only with a value of 0b.<br><br>Default value of this bit is 0b | RW |
| 11 | **NPEM Disabled Control -** When Set to 1b, this bit specifies that the NPEM Disabled indication be turned ON. This bit Cleared to 0b specifies that the NPEM Disabled indication be turned OFF.<br><br>If NPEM Disabled Capable bit in NPEM Capability register is 0b, this bit is permitted to be read-only with a value of 0b.<br><br>Default value of this bit is 0b | RW |
| 23:12 | **Reserved** | RsvdP |
| 31:24 | **Enclosure Specific -** The definition of enclosure specific bits is outside the scope of this specification.<br><br>Default value of this field is 00h | RW |

## 7.x.4. Native PCIe Enclosure Management Status Register (Offset 0Ch)

The NPEM Status Register contains the command completed bit to indicate to host software if command completed successfully.

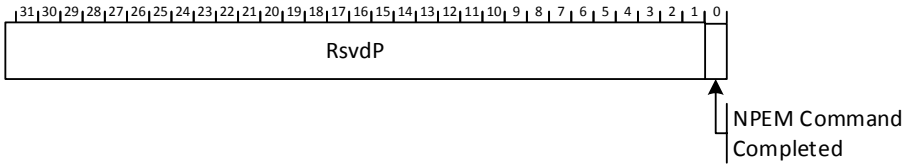


Figure 7-x4 NPEM Status Register

Table 7-x4 Native PCIe Enclosure Management Status Register

| Bit Location | Register Description | Attributes |
|---|---|---|

| | | |
|---|---|---|
| 0 | **NPEM Command Completed-**This bit is Set as an indication to host software that an NPEM command has completed successfully.<br><br>Default value of this bit is 0b.<br><br>Software must wait for an NPEM command to complete before issuing the next NPEM command. However, if this bit is not set within 1 second limit on command execution, software is permitted to repeat the NPEM command or issue the next NPEM command. If software issues a write before the Port has completed processing of the previous command and before the 1 second time limit has expired, the Port is permitted to either accept or discard the write. Such a write is considered a programming error, and could result in a discrepancy between the NPEM Control register and the enclosure element state. To recover from such a programming error and return the enclosure to a consistent state, software must issue a write to the NPEM Control register which conforms to the NPEM command completion rules. | RW1C |
| 31:1 | **Reserved** | RsvdZ |